2020

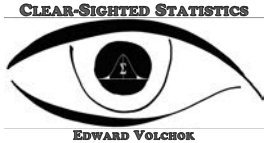
Clear-Sighted Statistics: Module 2: Types of Data (slides)

Edward Volchok
CUNY Queensborough Community College

Types of Data

Module 2

CLEAR-SIGHTED STATISTICS



EDWARD VOLCHOK

0

Lecture Objectives

Define Data	Distinguish Populations from Sample	Understand Sampling Error and Uncertainty
Understand the Fact-Value distinction	Identify the different types of data	Describe the four levels of measurement

1

What are Data?

2

What are data?

Data are a collection of numbers, words, pictures, measurements, observations, or descriptions of things

3

The word “data” is a plural noun; “datum” is singular

4

Data vs. Information

Data are raw, unorganized facts in need of processing

Information is data that has been organized in a context that makes it useful

In this class you’ll learn how to turn data into useful information

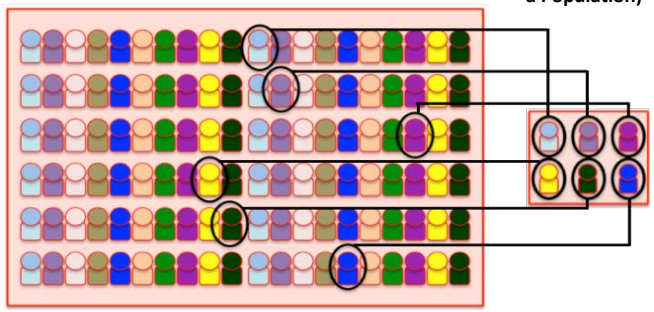
5

Distinguish Populations from Samples

6

Population
(All items of interest)

Sample
(Items selected from a Population)



7

Clear-Sighted Statistics

Parameters versus Statistics

Parameters

- Characteristics, traits, or variables of a population

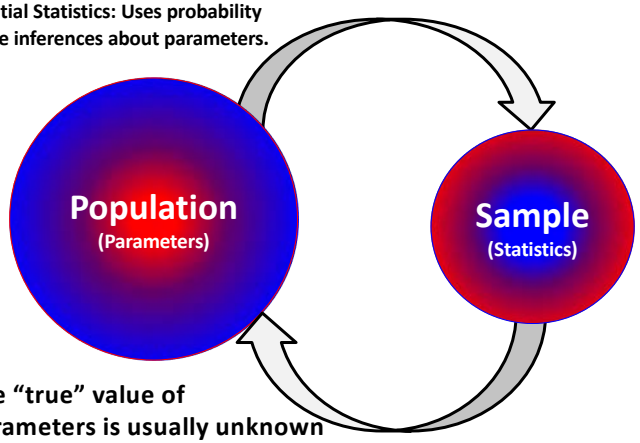
Statistics

- Characteristics, traits, or variables from a sample

8

Clear-Sighted Statistics

Inferential Statistics: Uses probability to make inferences about parameters.



The "true" value of parameters is usually unknown

9

Clear-Sighted Statistics

Inferential Statistics and Probability

Probability: A branch of mathematics that measures the likelihood that an event will occur

Probabilities are expressed as numbers between 0 and 1

The probability of an impossible event is 0 or 0%

An event that is certain to occur has a probability of 1 or 100%

10

Clear-Sighted Statistics

Inferential Statistics: Examples

Executives monitor the popularity of their brands by sampling customers and then making an inference about the attitudes of all customers

Chefs taste a few drops of their sauces to make decisions about the entire batch

Manufacturers producing a large volume of goods select a sample to determine if their output is meeting specifications

11

Clear-Sighted Statistics

Samples are used more often than populations to…

Save time

Save money

12

Clear-Sighted Statistics

Other reasons why samples are used

Samples sometimes destroy items of interest

Difficult to contact entire population

13

Clear-Sighted Statistics

When properly conducted, samples are *reliable*

14

Sample size is very important

Too large a sample wastes scarce resources

Too small a sample might fail to detect important effects or results

15

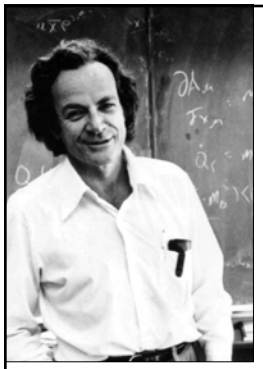
Sampling Error and Uncertainty

16

Sampling error is unavoidable

Sampling error is when the Statistic ≠ Parameter

17



Richard P. Feynman
Theoretical Physicist
(1918 - 1988)

"…Nature permits us to calculate only probabilities. Yet science has not collapsed."*

Richard P. Feynman. *QED: The Strange Theory of Light*. (Princeton, NJ, Princeton University Press, 1985), p. 19

18

Avoid Hasty Generalizations

- Generalizations are based on "anecdotal" evidence
- Two friends bought new iPhones that broke, therefore Apple is selling defective products
- Four students took 8 years to complete a bachelor's degree, therefore it must take longer to earn a college degree than 4 years
- My three-year-old nephew was vaccinated for measles and now he is autistic. Vaccinations cause autism.

19

Anecdotal evidence is faulty

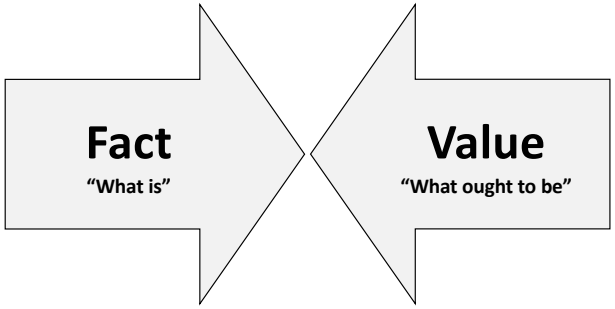
- Anecdotal evidence is based on data collected from a limited number of cases that may not represent the population
- Anecdotal evidence lacks scientific verification
- Data must be drawn from many cases in a manner that avoids bias

20

The Fact-Value distinction

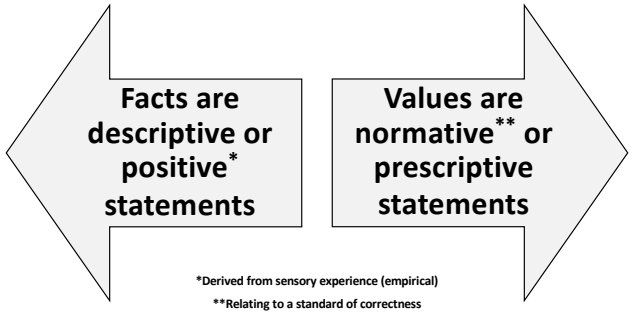
21

Fact-Value Distinction



22

Fact-Value Distinction



23

What are facts?

- Discovered using scientific methods
- Repeatable observations or measurements
- Also called empirical evidence
- Objective and verifiable

24

Facts about strawberries

- Among the first fruits to ripen in the spring
- Average strawberry has 200 seeds
- Average American eats 3.4 pounds annually (U.S. Dep't of Agriculture)
- California produces 72% of strawberries grown in the U.S.

25

What are values?

- Subjective
- Based on beliefs
- Cannot be verified empirically
- Can be culturally determined

26

Values about strawberries

- They are the best berries made by God
- Cherries taste better than strawberries

27

Values can be studied scientifically

Proportion of population that like strawberries	How strawberries rank on taste compared to other fruits
The intensity of people's feelings about strawberries	Cultural meanings associated with strawberries

28

Note: Post-Structuralist philosophers reject the Fact-Value Distinction

Jacques Derrida
(1930 - 2004)

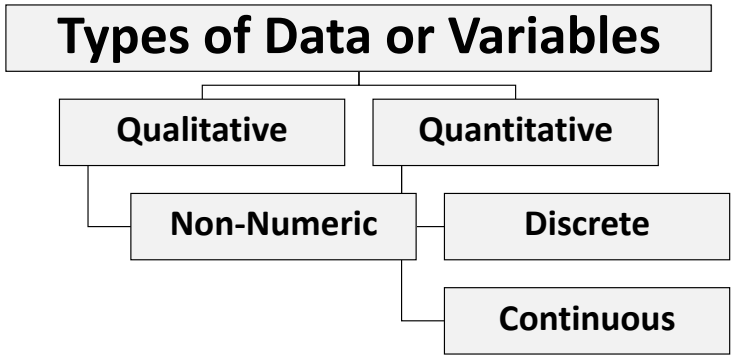
Michel Foucault
(1926 - 1984)

Bruno Latour
(1947 -)

29

The different types of data

30



The value of a "variable" varies by chance

31

Qualitative Variables

For a Qualitative, Categorical, or Attribute data, the characteristic being studied is non-numeric

Qualitative Data: Eye Color

Qualitative Data: Car Type

You Can Conduct Statistical Analysis With Qualitative Data

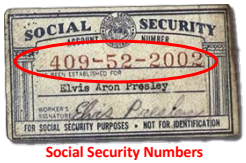
By counting the occurrences—frequencies—we can make quantitative data out of qualitative data

Percentages or Proportions (Relative Frequencies)

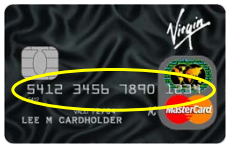
Quantitative Data

With a Quantitative Variable information is reported numerically

Beware:
Not all data with numbers are quantitative



Social Security Numbers



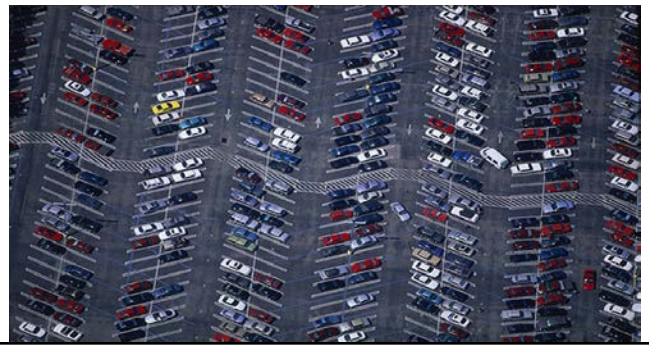
Credit Card Numbers

License Plate Numbers

Identification Numbers

Numbers are often qualitative identification codes

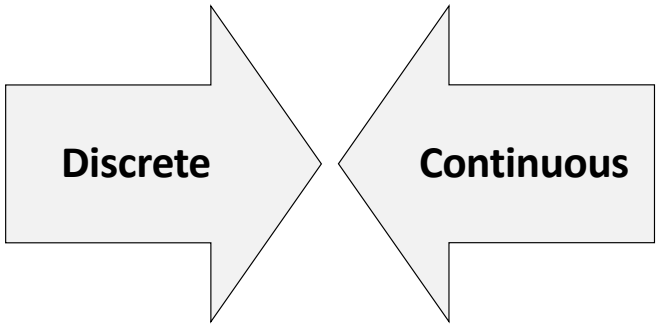
Quantitative Data:
Number of parking spots



Quantitative Data:
Number of dogs in these pictures

Animal	f	%
Dogs	62	98.41%
Cats	1	1.59%
Total	63	100.00%

Quantitative variables can be…



40

Discrete Variables

Discrete Variables can only assume certain values and there are usually "gaps" between values

Typically values are whole numbers or integers, which can be counted: *1, 2, 3, … n*

41

Quantitative Data: Number of kids in this picture

20

42

Discrete Variables

Number of rooms in this house

43

Continuous Variables

A Continuous Variable can assume any value within a specified range

44

Continuous Variable: Time to complete a race

Time to run a mile/1600 m

World Record
3 minutes, 43.13 seconds

Hicham El Guerroij
Morocco
3:43:13
7/7/99

https://www.youtube.com/watch?v=Ji0vK7fV5Rk

45

Continuous Variable: Air Pressure

46

Continuous Variable: Weight

47

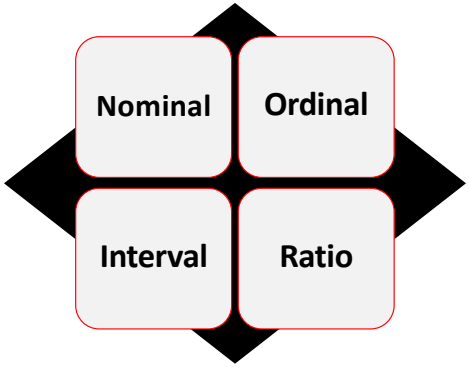
Continuous Variable: Height

48

The 4 levels of measurement

49

Levels of Measurement (NOIR)



50

Levels of Measurement: Qualitative & Quantitative Variables

Qualitative Variables:	Quantitative Variables	Why is this important?
• Also known as Nominal & Ordinal Levels of Measurement	• Also known as Interval & Ratio Levels of Measurement	• Techniques used to display, summarize, and analyze data depend on whether variables are qualitative or quantitative

51

Clear-Sighted Statistics

Nominal Level

Qualitative or Categorical Data

Data classified into categories that cannot be arranged in a meaningful order

Can add a quantitative dimensions by counting frequencies and calculating proportions

52

Clear-Sighted Statistics

Nominal Level: Color of your socks

53

Clear-Sighted Statistics

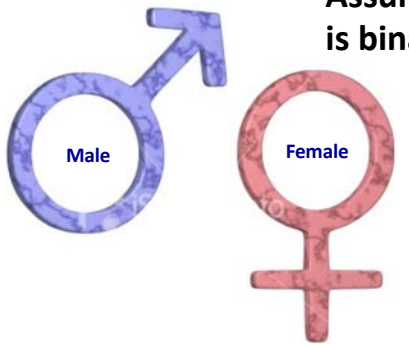
Nominal Level: Color of Yoshi the Cat

54

Clear-Sighted Statistics

Nominal Level: Gender

Assuming gender is binary



55

Nominal Level: Numbers on Athletes' Uniforms

56

Nominal Level: Religious Affiliation

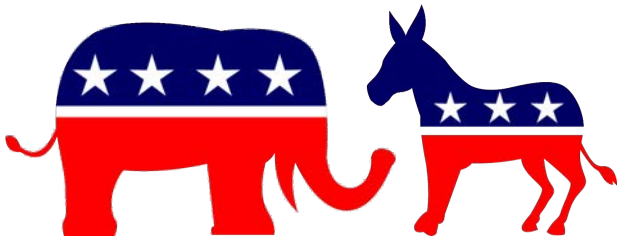
57

Nominal Level: Birthplace

Beijing
New York City
Santo Domingo
Dhaka
San Juan
Mumbai
Baghdad
Shanghai
Islamabad
Moscow
Warsaw

58

Nominal Level: Political Party



59

Nominal Level data must be…

Mutually Exclusive	Collectively Exhaustive
• A measurement or object is included in only one category	• Each measurement or object must appear in one of the categories

60

Ordinal Level (Qualitative Data)

Data can be ordered

Differences between data values cannot be determined, or are meaningless

61

Ordinal Level: The Hotness of Chili Peppers

62

Ordinal Level: Order of finish

3rd 1st 2nd

63

Ordinal Level: These tee shirt sizes

64

Letter Grades

A = Excellent

B = Good

C = Satisfactory

D = Passing

F = Failing

65

Ordinal Level: Likert Questions*

(Despite what some researchers think!)

Donald Trump is the best president since Lincoln.

Strongly Disagree	Disagree	Neither Agree Not Disagree	Agree	Strongly Agree

*Widely used psychometric scale developed by Rensis Likert

66

Interval Level (Quantitative Data)

Like the ordinal scale, with quantitative differences between data values

No natural zero

No true nothingness!

67

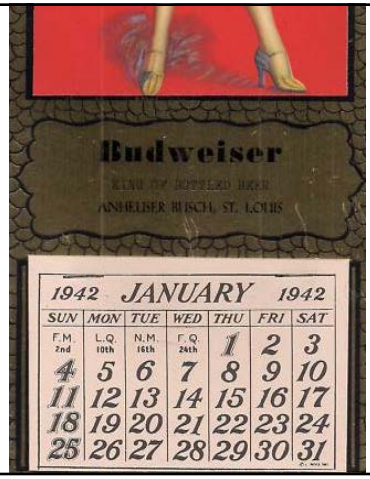
Interval Level: Temperature

Fahrenheit and Celsius scales

68

Interval Level: Calendar Dates

There's no Day Zero!



69

Interval Level: Calendar Years

There's no Year Zero!

Time did not start in 1 CE

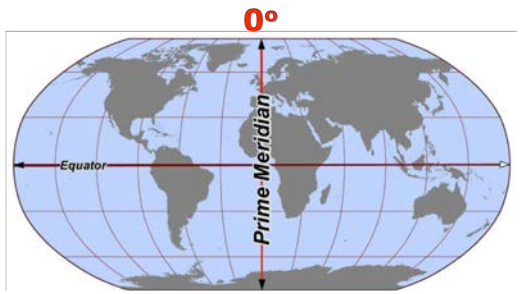
70

Interval Level: Tide Heights

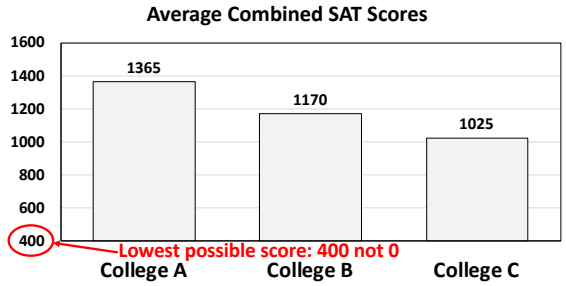
71

Interval Level: Map Longitudes

The Prime Meridian is an arbitrary zero



Interval Level: SAT Scores



Ratio Level (Quantitative Data)

Like a interval level

But with a non-arbitrary zero

Ratio Level: Scoville Heat Index

Ratio Level: Time remaining in class

76

Ratio Level:

Money in your pocket

77

Ratio Level: Money in Jeff Bezo's Wallet

78

Ratio Level: Course credits this semester

Semester 4		
BU-401 or BU-701	Principles of Marketing or Principles of Finance	3
_____	Foreign Language and/or Liberal Arts & Sciences*	0-4
HI-110, 111, or 112	History	3
BU-512	Introduction to Information Systems and Technologies	3
_____	Free Electives**✥✥	1-6
	Sub-total	14-15
Total Credits Required		**60**

79

Clear-Sighted Statistics

Levels of Measurement: Summary

	Categorical Data		Quantitative Data	
	Nominal	Ordinal	Interval	Ratio
Names	X	X	X	X
Order		X	X	X
Intervals			X	X
Absolute Zero				X
Example:	Gender, Hair Color, Zip Code. Make of Car	Military rank, Order of finish (1^{st}, 2^{nd}, 3^{rd})	Shoe Size, Temperature, SAT Scores,	Income, Weight, Height, Distance Travelled

80

Clear-Sighted Statistics

Level	Definition	Statistics
Nominal	Qualitative data, numbers are just names, no ordering possible. **Examples:** Eye color, Student ID numbers, Social Security Numbers, your birthplace, dog breeds.	• Frequencies • Mode • Chi-Square
Ordinal	Qualitative data that can be counted and ranked. **Examples:** Order of finish in a race (1^{st}, 2^{nd}, 3^{rd}), Likert responses (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).	• Frequencies • Mode • Median • Percentiles • Range and Inter-Quartile Range • Chi-Square
Interval	Quantitative Data: Known distance between ranks. But, no absolute zero. Most arithmetic operations. **Examples:** Temperature measured on the Fahrenheit or Celsius scales, SAT scores.	• Frequencies • Mode • Median • Mean • Standard Deviation • Parametric Hypothesis tests • Correlation & Regression
Ratio	Quantitative Data: Scale possesses an absolute zero. All arithmetic operations. **Examples:** Height, weight, age, income, temperature on the Kelvin scale.	• Frequencies • Mode • Median • Mean (Including the Geometric & Harmonic means) • Standard Deviation • Parametric Hypothesis tests • Correlation & Regression

81

Clear-Sighted Statistics

Levels of Measurement

Ratio:
Values are ranked, distance between each values is known, and there is an absolute zero

Interval:
Values are ranked with a known distance between ranks, but there is no absolute zero

Quantitative

Ordinal:
Values can be ranked: 1^{st}, 2^{nd}, 3^{rd}, distance between ranks unknown

Qualitative

Nominal:
Values are qualitative categories

82

83

CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK

84